US007415416B2

(12) **United States Patent**
Rees

(10) **Patent No.: US 7,415,416 B2**
(45) **Date of Patent: Aug. 19, 2008**

(54) **VOICE ACTIVATED DEVICE**

(75) Inventor: **David Llewellyn Rees**, Bracknell (GB)

(73) Assignee: **Canon Kabushiki Kaisha**, Tokyo (JP)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 699 days.

(21) Appl. No.: **10/937,558**

(22) Filed: **Sep. 10, 2004**

(65) **Prior Publication Data**

US 2005/0102133 A1 May 12, 2005

(30) **Foreign Application Priority Data**

Sep. 12, 2003 (GB) 0321447.5

(51) **Int. Cl.**
G10L 15/00 (2006.01)

(52) **U.S. Cl.** **704/275**; 704/270; 704/233; 704/214; 704/217

(58) **Field of Classification Search** 704/270, 704/275, 233, 214, 217
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,951,079 A		8/1990	Hoshino et al.	
5,027,149 A	*	6/1991	Hoshino et al.	396/56
5,983,186 A	*	11/1999	Miyazawa et al.	704/275
6,049,766 A		4/2000	Laroche	
6,272,460 B1	*	8/2001	Wu et al.	704/226
6,711,536 B2		3/2004	Rees	
2003/0154078 A1		8/2003	Rees	
2003/0163313 A1		8/2003	Rees	

FOREIGN PATENT DOCUMENTS

EP	1 187 099	3/2002
JP	60-205432	10/1985
JP	60-205433	10/1985
JP	2000-227633	8/2000
JP	2001-305642	11/2001
JP	2002-354335	12/2002

OTHER PUBLICATIONS

R.Tucker□□"Voice activity detection using a periodicity measure"□□IEEE, Aug. 1992, pp. 377-380.*

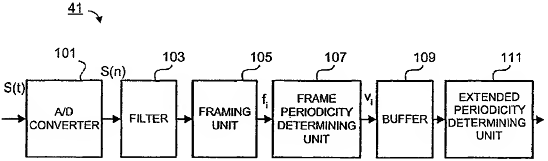
* cited by examiner

Primary Examiner—Daniel D Abebe
(74) *Attorney, Agent, or Firm*—Fitzpatrick, Cella, Harper & Scinto

(57) **ABSTRACT**

A voice activated camera is described which allows users to take remote photographs by speaking one or more keywords. In a preferred embodiment, a speech processing unit is provided which is arranged to detect extended periodic signals from a microphone of the camera. A control unit is also provided to control the taking of a photograph when such an extended periodic component is detected by the speech processing unit.

15 Claims, 6 Drawing Sheets



shown in FIG. 2, the camera 3 also includes a microphone 39 for converting a user's speech into corresponding electrical speech signals; and a speech processing unit 41 which processes the electrical speech signals to detect the presence of a keyword in the user's speech and which informs the camera control unit 33 accordingly.

Speech Processsing Unit

As discussed above, the speech processing unit 41 is arranged to detect keywords spoken by the user in order to control the taking of remote photographs. In this embodiment, the speech processing unit does not employ a "conventional" automatic speech recognition type keyword spotter which compares the spoken speech with stored models to identify the presence of one of the keywords. Instead, the speech processing unit 41 used in this embodiment is arranged to detect a sustained periodic signal within the input speech, such as would occur if the user says the word "cheeeese" or some other similar word. The inventor has found that because of the strong periodic nature of such a sustained vowel sound, the speech processing unit 41 can still detect the sound even at very low signal-to-noise ratios.

The way in which the speech processing unit 41 operates in this embodiment will now be explained with reference to FIGS. 3 to 7.

FIG. 3 illustrates the main functional blocks of the speech processing unit 41 used in this embodiment. The input signal (S(t)) received from the microphone 39 is sampled (at a rate of just over 11 KHz) and digitised by an analogue-to-digital (A/D) converter 101. Although not shown, the speech processing unit 41 will also include an anti-aliasing filter before the A/D converter 101, to prevent aliasing effects occurring due to the sampling. The sampled signal is then filtered by a bandpass filter 103 which removes unwanted frequency components. Since voiced sounds (as opposed to fricative sounds) are generated by the vibration of the user's vocal cords, the smallest fundamental frequency (pitch) of the periodic signal to be detected will be approximately 100 Hertz Therefore, in this embodiment, the bandpass filter 103 is arranged to remove frequency components below 100 Hertz which will not contribute to the desired periodic signal. Also, the bandpass filter 103 is arranged to remove frequencies above 500 Hertz which reduces broadband noise from the signal and therefore improves the signal-to-noise ratio. The input speech is then divided into non-overlapping equal length frames of speech samples by a framing unit 105. In particular, in this embodiment the framing unit 105 extracts a frame of speech samples every 23 milliseconds. With the sampling rate used in this embodiment, this results in each frame having 256 speech samples. FIG. 4 illustrates the sampled speech signal (S(n), shown as a continuous signal for ease of illustration) and the way that the speech signal is divided into non-overlapping frames.

As shown in FIG. 3, each frame f_1 of speech samples is then processed by a frame periodicity determining unit 107 which processes the speech samples within the frame to calculate a measure (v_i) of the degree of periodicity of the speech within the frame. A high degree of periodicity within a frame is indicative of a voiced sound when the vocal cords are vibrating. A low degree of periodicity is indicative of noise or fricative sounds. The calculated periodicity measure (v_i) is then stored in a first-in-first-out buffer 109. In this embodiment, the buffer 109 can store frame periodicity measures for forty-four consecutive frames, corresponding to just over one second of speech. Each time a new frame periodicity measure is added to the buffer 109, an extended periodicity determining unit 111 processes all of the forty-four periodicity measures in the buffer 109 to determine whether or not a sustained periodic sound is present within the detection window represented by the forty-four frames.

When the extended periodicity determining unit 111 detects a sustained periodic sound within the speech signal, it passes a signal to the camera control unit 33 confirming the detection. As discussed above, the camera control unit 33 then controls the operation of the camera 3 to take the photograph at the appropriate time.

Frame Periodicity Determining Unit

As those skilled in the art will appreciate, various techniques can be used to determine a measure of the periodicity of the speech within each speech frame. However, the main components of the particular frame periodicity determining unit 107 used in this embodiment is shown in FIG. 5. As shown, the frame periodicity determining unit 107 includes an auto-correlation determining unit 1071 which receives the current speech frame f_i from the framing unit 105 and which determines the auto-correlation of the speech samples within the frame. In particular, the auto-correlation determining unit 1071 calculates the following function:

$$A(L) = \frac{1}{N-L}\sum_{j=0}^{N-L-1} x(j)x(j+L) \tag{1}$$

where x(j) is the j^{th} sample within the current frame, N is the number of samples in the frame, j=0 to N−1 and L=0 to N−1.

The value of A(L) for L=0 is equal to the signal energy and for L>0 it corresponds to shifting the signal by L samples and correlating it with the original signal. A periodic signal shows strong peaks in the auto-correlation function for values of L that are multiples of the pitch period. In contrast, non-periodic signals do not have strong peaks.

FIG. 6 shows the auto-correlation function ($A_i(L)$) for a frame of speech f_i representing a speech signal which is periodic and which repeats approximately every 90 samples. As shown in FIG. 6, the auto-correlation around L=180. Further, the value of the auto-correlation function at L=90 is approximately the same as the value at L=0, indicating that the signal is strongly periodic.

The fundamental frequency or pitch of voiced speech signals varies between 100 and 300 Hertz. Therefore, a peak in the auto-correlation function is expected between $L_{low}=F_s/300$ and $L_{high}=F_s/100$, where F_s is the sampling frequency of the input speech signal. Consequently, in this embodiment, the auto-correlation function output by the auto-correlation determining unit 1071 is input to a peak determining unit 1073 which processes the auto-correlation values between $A(L_{LOW})$ and $A(L_{HIGH})$ to identify the peak value $(A(L_{MAX}))$ within this range. In this embodiment, with a sampling rate of just over 11 kHz the value of L_{LOW} is 37 and the value of L_{HIGH} is 111. This search range of the peak determining unit 1073 is illustrated in FIG. 6 by the vertical dashed lines, which also shows the peak occurring at $L_{MAX}=90$. The auto-correlation values A(0) and $A(L_{MAX})$ are then passed from the peak determining unit 1073 to a periodicity measuring unit 1075 which is arranged to generate a normalised frame periodicity measure for the current frame (f_i) by calculating:

$$v_i = \frac{A_i(L_{MAX})}{A_i(0)} \tag{2}$$

where v_i will be approximately one for a periodic signal and close to zero for a non-periodic signal.